STAT 672 Statistical Learning & Data Analytics Spring 2018
George Mason University
Instructor: Martin Slawski

Homework 1

Pencil-and-paper part due Thursday, Feb 15, in class. Code and relevant output has to be submitted via Blackboard. Please follow the following format for the filename of your submission: `Lastname_Firstname_HWx.zip`, where `x` needs to be substituted by the homework #.
Maximum possible points to earn: 18; 100% = 15 points, rest counts as bonus.
!Copying from the solution of the previous year is regarded as an honor code violation.!

Problem 1 (9 points)

Drawing random vectors from the unit ball $B_2^d := \{x \in \mathbb{R}^d : \ \|x\|_2 \leq 1\}$.

a)
One method to draw random vectors from B_2^d is *rejection sampling* by sampling from the unit cube $B_\infty^d := \{x \in \mathbb{R}^d : \ \|x\|_\infty \leq 1\}$ and accepting only those samples with Euclidean norm no larger than one, that is:

1. Draw $X \sim B_\infty^d$ (can be done by sampling each coordinate X_j uniformly from $[-1, 1]$).
2. Keep X if $\|X\|_2 \leq 1$; otherwise, discard X.

Repeat 1. and 2. until the desired number of samples has been generated.

Using Stirling's approximation, compute the acceptance probability

$$\mathbb{P}_{X \sim B_\infty^d}(\|X\|_2 \leq 1) = \frac{\text{vol}(B_2^d)}{\text{vol}(B_\infty^d)}$$

asymptotically as d gets large. *Hint*. The volume of B_2^d is given by

$$\text{vol}(B_2^d) = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)},$$

where Γ denotes the Gamma function. What does the result imply for the expected runtime to generate a single random vector from B_2^d with rejection sampling?

b)
A more efficient approach than in a) is based on spherical coordinates. We generate X as $X = \Theta \cdot R$ with Θ being a random vector from the uniform distribution on the unit sphere in $\mathbb{R}^d$ and $0 \leq R \leq 1$ being the random radius from an appropriate distribution.

1. We can generate Θ as $\Theta = Z/\|Z\|_2$, where $Z \sim N(0, I_d)$. Explain why.
2. As mentioned in class, $\mathbb{P}_{X \sim B_2^d}(\|X\|_2 \leq t) = t^d$ for $0 \leq t \leq 1$. Based on this result,

we can generate $\|X\|_2 = R \sim U^{1/d}$, where U is uniform on $[0,1]$. Explain why.
Hint. Read about *inversion sampling*.

c)
Compare the runtime for generating 100 samples according to the methods in a) and b) for $d = 10$ empirically.

Problem 2 (9 points)

a)

Generate two random vectors $X_+ \sim N(\mu_+, 4 \cdot I_d)$ and $X_- \sim N(\mu_-, I_d)$, where $\mu_\pm = (\pm 5/\sqrt{d})\mathbf{1}_d$[1]. Then generate a random vector Z from the same distribution as X_+.

Repeat 1000 times and compute the relative frequency of the event

error $= \{\|X_+ - Z\|_2 \geq \|X_- - Z\|_2\}$

for $d \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$. Describe the results. Are they intuitive/counter-intuitive? Explain why.

Hint. Sampling from a d-dimensional Normal distribution with diagonal covariance matrix is easy: the components are then independent, and can be generated one by one from a one-dimensional Normal distribution.

b)

Compute the expected squared distances $\mathbb{E}[\|X_+ - Z\|_2^2]$ and $\mathbb{E}[\|X_- - Z\|_2^2]$. Relate the result to a).

Hint. Expand $\|v\|_2^2 = \sum_{j=1}^d v_j^2$ and use linearity of expectation. Use that X_+ and Z respectively X_- and Z are independent.

[1] $\mathbf{1}_d$ is a d-dimensional vector of all ones.